DOCUMENT RESUME

ED 167 622 TM 008 436

AUTHOR Yap, Kim Onn
TITLE Can Selection Tests Be Used As Pretest?
PUB DATE Mar 78
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Achievement Gains; *Analysis of Variance; Elementary Secondary Education; Low Achievers; Predictor Variables; *Pretests; *Scoring Formulas; *Statistical Bias; *Test Reliability; *True Scores
IDENTIFIERS Elementary Secondary Education Act Title I; RMC Models

ABSTRACT

A simulation study was designed to assess the severity of regression effects when a set of selection scores is also used as pretest scores as this pertains to RMC Model A of the Elementary and Secondary Education Act Title I evaluation and reporting system. Data sets were created with various characteristics (varying data reliability and extremeness of subgroups) that are relevant to the regression phenomenon. These data sets were analyzed to obtain indices of the amount of regression which might occur under various conditions. An adjustment method was presented which predicted pretest scores on the basis of selection scores and the correlation between selection and pretest, the correlation coefficient being replaceable by such indices of reliability as the internal consistency coefficient. Results suggest that when data reliability is high (.94 or above) the impact of regression effects appears small; but when reliability is low the effects of regression are hard to predict. When selection test scores are of high reliability, such regressed scores may be used as a pretest measure without severely biasing the evaluation results. Results can be used to formulate a rule-of-thumb for adjusting selection scores when they are to be used as pretest scores also. (Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED167622

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

CAN SELECTION TESTS BE USED AS PRETESTS?

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Kim Onn Yap

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Kim Onn Yap

Northwest Regional Educational Laboratory

A paper presented as part of a symposium,
Title I Evaluation: An Examination of Model A,
Annual Meeting of the American Educational Research Association
Toronto, Canada, March 27-31, 1978
Session #29.07

TM008 436

CAN SELECTION TESTS BE USED AS PRETESTS?

Kim Onn Yap

Northwest Regional Educational Laboratory

INTRODUCTION

Model A of the proposed ESEA Title I evaluation and reporting system (Tallmadge & Wood, 1976; Tallmadge, 1976) specifies that independent measures should be used for selection and pretest to eliminate regression effects due to measurement error. These effects are expected to occur when extreme groups (e.g., the bottom 20 percent of the school population) are selected to participate in Title I programs. The size of regression is a function of the correlation between the two measures involved, i.e., the correlation between selection and pretest. A second parameter which affects the size of regression is the difference between the subgroup means. That is, the greater the mean difference the more regression (Campbell and Erlehacher, 1970).

The observed score is an unbiased estimate of the true score only for unselected scores, that is, for total sets of scores that have been allowed to fall freely. Where scores have been selected because of their observed values, they become biased estimates of true scores in the direction indicated by regression toward the population mean.

More specifically, the regression phenomenon stems from the behavior of measurement error. Positive measurement error contributes predominantly to above-average test scores; negative measurement error contributes predominantly to below-average test scores. A higher score is more likely to contain a positive error than a lower score, and vice versa. While error can be expected to be randomly distributed

across scores for the total group, its means are necessarily not equal to zero for low and high scoring subgroups.

The regression effect is removed when the low-scoring subgroup is retested, error being again randomly distributed across the group members. Negative error and positive error cancel each other, yielding an expected mean error of zero. This results in retest scores that are higher than the first set of scores which contained a preponderance of negative error. The higher retest scores are closer to the mean of the total group and also closer to the true scores.

The regression effect poses a problem for Model A in that the difference between the retest scores and the first set of test scores could mistakenly be interpreted as an achievement gain--when no real difference in achievement exists. It follows that if the first set of test scores is used as a pretest measure and the second a posttest measure, we will obtain a biased estimate of achievement gain. It further follows that if the first set of scores is used for selection purposes, it will be necessary to retest the selected subgroups to get an unbiased pretest measure. A third set of test scores will need to be collected to provide an unbiased posttest measure.

THE PROBLEM AND PURPOSE OF THE STUDY

In Title I schools, low-scoring students are typically selected for participation in Title I projects. This is an appropriate procedure for identifying low-achieving students--although some students not selected may have lower true achievement levels than some who are selected (Tallmadge, 1976). The observed scores of the selected students will, however, contain a preponderance of negative measurement error. If used as pretest scores, these observed scores will provide a biased estimate of the students' initial achievement level and, therefore, represent a biased basis for estimating achievement gain at posttest. The solution, as indicated earlier, is to obtain a second set of test scores to serve as the pretest measure.

Unfortunately, for a number of reasons, this solution is not always available to Title I project administrators. Such factors as school schedule, budget and time constraints (not to mention the burden of overtesting students) often make it impractical or impossible to obtain a second set of test scores as the pretest measure. In other words, the same set of test scores will necessarily have to be used as the selection criterion as well as the pretest measure.

This points out a need to go beyond explaining that regression bias exists in those situations. We need to estimate the amount or severity of the bias when the selection test is also used as the pretest. In addition, there is a need to assess the performance of any statistical procedures that may exist for reducing or eliminating the bias.

The purpose of this study was therefore twofold. First, the investigation was an attempt to estimate the severity of regression bias when a selection measure is also used as the pretest measure. Secondly, the study was aimed at assessing the performance of existing procedures for reducing or eliminating the bias. The study essentially consisted of a simulation of data sets resembling those typically encountered in Title I evaluation and an analysis of the simulated data. It was hoped that some practical guidelines could be developed which would suggest conditions under which a selection measure may also be used as the pretest measure.

THE SIMULATION

Data sets were generated by using the following formulae which defined selection and pretest scores:

$$X_1ij = Tij + Eij, \text{ and}$$

$$X_2ij = Tij + E'ij,$$

where X_1ij is the selection score of person j in group i, X_2ij is the pretest score of person j in group i, Tij is the true score of person j in group i, and Eij and E'ij are error scores at selection and pretest times, respectively, of person j in group i, (i = 1, ... , 100; j= 1, ... , 100).

A computer program was developed to generate values for the simulated selection and pretest scores. These values were derived from random numbers provided by subroutine GAUSS (IBM, 1968). For example, a pair of selection and pretest scores may be obtained as follows:

$$Tij = .9\ (N_1),$$

$$Eij = .1\ (N_2), \text{ and}$$

$$E'ij = .1\ (N_3),$$

where the N's are two-digit random numbers provided by GAUSS.

Subroutine GAUSS generates normally distributed random numbers with a given mean and standard deviation which were chosen to be 50 and 21.06, respectively, for this study. They are also the mean and standard deviation of the normal curve equivalent (NCE) scores.

It should be obvious that by varying the multipliers of the random numbers one could manipulate the reliability of the data. In the above example we have set the true score (Tij) and error score (Eij) to be nine-tenths and one-tenth, respectively, of the observed selection score (X_1ij). The same is true of the observed pretest score. In other words, the ratio of the true score to the error score in terms of size is 9:1 for this pair of selection and pretest scores. If we had set Tij = .8 (N_1), Eij = .2 (N_2) and E' ij = .2 (N_3) we would have had larger errors and less reliability for X_1ij and X_2ij. The size of the true score relative to the error score was described as the error ratio in this study.

Four error ratios were used in generating the data sets: 9:1, 8:2, 7:3 and 6:4. This means that data of varying levels of reliability were simulated. More specifically, 100 data sets (each consisting of 100 simulated cases) were created for each of the four error ratios. The entire simulation thus included 400 data sets with a total of 40,000 hypothetical cases. The sample means of these data sets range from 49.86 to 50.51, closely approximating the population mean of 50. The standard deviations, however, tend to be generally smaller than the population standard deviation of 21.06, ranging from 15.26 to 19.03. In all cases, the selection and related pretest scores are shown to have approximately the same mean and standard deviation. Sample means and standard deviations of the data sets are reported in Table 1.

TABLE I

Overall Means and Standard Deviations of the Simulated Data Sets*

Error Ratio	Mean		Standard Deviation	
	Selection	Pretest	Selection	Pretest
9:1	50.51	50.49	19.03	19.02
8:2	50.14	50.10	17.41	17.38
7:3	49.86	49.86	15.92	16.05
6:4	49.96	50.04	15.26	15.27

*These means and standard deviations were each based on 100 data sets each consisting of 100 simulated cases.

A word should be said about the relationship between what was described as the error ratio and reliability coeffient as the term is used in the measurement literature. As defined by Gulliksen (1950), the reliability coefficient is the ratio of true variance to observed variance. Since multiplying an element by a constant will multiply the variance by the square of the constant, the ratio of true variance to error variance will be greater than the error ratio. Thus, error ratios should be considered as lower bound reliability coefficients. That is to say, an error ratio of 9:1, for example, represents a reliability coefficient which is in fact higher than .90.

As indicated earlier, the amount of regression is a direct function of the correlation coefficient between selection and pretest. In standard score from (i.e., as deviations from the means divided by their standard deviations) one can predict a pretest score on the basis of the selection score and the correlation coefficient between selection and pretest as follows:

$$\hat{x}_2 = rx_1, \qquad \text{where}$$

x_1 is a selection score, $\hat{x}_2$ indicates a predicted (or adjusted) pretest score, and r is the correlation coefficient between selection and pretest for the unselected group. Making certain assumptions about the means and variances of the selection and pretest scores (i.e., both have the same mean and variance for the unselected group) it can be shown, in raw score form, that

$$\hat{X}_2 = \overline{X}_1 + r(X_1 - \overline{X}_1), \qquad \text{where}$$

X_1 is a selection raw score, $\overline{X}_1$ is the selection mean score, $\hat{X}_2$ is a predicted pretest score, and r is the correlation coefficient between selection and pretest scores for the unselected group.

The present study used this adjustment method to predict pretest scores and assessed its performance with data of various reliability levels.

ANALYSIS AND RESULTS

Two types of analyses were performed on the data. First, mean differences between selection and pretest scores were computed for each of the data sets. These mean differences were then summed and averaged for each of the data types (i.e., data sets of different levels of reliability) as an overall index of the regression effect. Standard deviations were computed to indicate the distributions of these mean differences. Secondly, mean differences between pretest scores and predicted pretest scores were computed for each of the data sets. Again, these mean differences were summed and averaged for each of the data types as an overall index of the adequacy of the adjustment method. Standard deviations were computed to indicate the distributions of the mean differences.

The analyses were performed for the total sample and three subsamples representing extreme subgroups. These subsamples consisted of the bottom 30, 20 and 10 percent of the cases in the total sample. Results of the analyses are presented in Tables 2 and 3.

TABLE 2

Mean Differences Between Selection and Pretest*

Sample	Error Ratio**	Mean Difference	Standard Deviation of Difference
Total	9:1 (r = .99)	.02	.27
	8:2 (r = .94)	.04	.57
	7:3 (r = .84)	-.00	.96
	6:4 (r = .70)	-.07	1.30
Bottom 30%	9:1	-.24	.58
	8:2	-1.23	1.07
	7:3	-2.66	1.64
	6:4	-5.51	2.21
Bottom 20%	9:1	-.32	.70
	8:2	-1.53	1.47
	7:3	-3.25	1.91
	6:4	-6.37	2.61
Bottom 10%	9:1	-.38	.97
	8:2	-1.72	2.04
	7:3	-4.35	2.68
	6:4	-8.06	3.27

*Each mean difference was based on 100 data sets.

**Error ratios should be considered as lower bound reliability. In other words, an error ratio of 9:1 represents a reliability coefficient higher than .90. The average correlation coefficients between selection and pretest for the total samples are in parentheses.

TABLE 3

Mean Differences Between Pretest and Predicted Pretest*

Sample	Error Ratio**	Mean Difference	Standard Deviation of Difference
Total	9:1 (r = .99)	-.02	.27
	8:2 (r = .94)	-.04	.57
	7:3 (r = .84)	.00	.96
	6:4 (r = .70)	.07	1.30
Bottom 30%	9:1	-.03	.57
	8:2	.05	1.03
	7:3	-.20	1.51
	6:4	.16	2.03
Bottom 20%	9:1	-.00	.69
	8:2	.12	1.43
	7:3	-.20	1.75
	6:4	-.10	2.30
Bottom 10%	9:1	-.02	.95
	8:2	-.04	2.00
	7:3	.06	2.44
	6:4	-.02	2.97

*Each mean difference was based on 100 data sets.

**Error ratios should be considered as lower bound reliability. In other words, an error ratio of 9:1 represents a reliability coefficient higher than .90. The average correlation coefficients between selection and pretest for the total samples are in parentheses.

The same analyses were repeated on the absolute mean differences for each of the data sets. These analyses were performed to assess the magnitude of mean differences, regardless of the direction (positive or negative) of the differences, between selection and pretest scores and between pretest and predicted pretest scores. As would be expected, random fluctuations yielded a number of instances in which the selection mean score was actually higher than the pretest mean score. For the same reason, the predicted pretest mean score was higher than the pretest mean score in some cases and lower in others. Results of the analyses on absolute mean differences were summarized in Tables 4 and 5.

TABLE 4

Mean Differences Between Selection and Pretest Based on Absolute Values*

Sample	Error Ratio**	Mean Difference	Standard Deviation of Difference
Total	9:1 (r = .99)	.22	.16
	8:2 (r = .94)	.47	.32
	7:3 (r = .84)	.75	.60
	6:4 (r = .70)	1.07	.74
Bottom 30%	9:1	.51	.36
	8:2	1.35	.92
	7:3	2.74	1.51
	6:4	5.54	2.12
Bottom 20%	9:1	.62	.46
	8:2	1.73	1.23
	7:3	3.27	1.87
	6:4	6.38	2.58
Bottom 10%	9:1	.85	.59
	8:2	2.20	1.50
	7:3	4.46	2.50
	6:4	8.06	3.27

*Each mean difference was based on 100 data sets.

**Error ratios should be considered as lower bound reliability. In other words, an error ratio of 9:1 represents a reliability coefficient higher than .90. The average correlation coefficients between selection and pretest for the total samples are in parentheses.

TABLE 5

Mean Differences Between Pretest and Predicted Pretest Based on Absolute Values*

Sample	Error Ratio**	Mean Difference	Standard Deviation of Difference
Total	9:1 (r = .99)	.22	.16
	8:2 (r = .94)	.47	.32
	7:3 (r = .84)	.75	.60
	6:4 (r = .70)	1.07	.74
Bottom 30%	9:1	.45	.35
	8:2	.83	.61
	7:3	1.21	.92
	6:4	1.61	1.24
Bottom 20%	9:1	.54	.42
	8:2	1.15	.84
	7:3	1.42	1.02
	6:4	1.78	1.44
Bottom 10%	9:1	.78	.55
	8:2	1.56	1.24
	7:3	1.95	1.47
	6:4	2.36	1.78

*Each mean difference was based on 100 data sets.

**Error ratios should be considered as lower bound reliability. In other words, an error ratio of 9:1 represents a reliability coefficient higher than .90. The average correlation coefficients between selection and pretest for the total samples are in parentheses.

In looking at the results summarized in Tables 2 to 5, it is clear that, overall, the regression bias does exist. It is also clear that the size of such bias is a direct function of data reliability and the "extremeness" of the subgroup. Thus, relatively small regression effects were found with respect to data sets of high levels of reliability (e.g., data sets with error ratios of 9:1 and 8:2). As shown in Table 2, the mean differences range from .04 to -1.72 for the total sample and all subsamples. (The mean differences reported for the total sample may best be regarded as a result of normal random fluctuations rather than regression.) As the level of reliability drops (e.g., in data sets with error ratios of 7:3 and 6:4) the impact of regression increases, with the mean differences between selection and pretest ranging from -2.66 to -8.06 for the subsamples.

It is interesting to note that as the error ratio drops from one level to the next, the regression effects tend to double in size. Take the bottom 30 percent subgroup for example. When the error ratio drops from 8:2 to 7:3, the mean difference increases in size from -1.23 to -2.66. Similarly, when the error ratio declines from 7:3 to 6:4, the mean difference increases from -2.66 to -5.51. This pattern of results is observed in all the subsamples.

The standard deviations for the mean differences appear to be relatively large, indicating great dispersion in the distributions. Confidence intervals for the mean differences are thus fairly wide which suggests caution in using any guidelines which might be formulated on the basis of these results.

As described earlier, predicted pretest scores were obtained on the basis of selection scores and the correlation coefficient between selection and pretest for the unselected group (i.e., the total sample). Results in Table 3 suggest that this procedure

works remarkably well with all the data sets. The mean differences between predicted and actual pretest scores range from -.20 to .16 for all the subsamples. The standard deviations, however, appear relatively large, suggesting substantial random fluctuations in the mean differences.

It should be noted that the correlation coefficient between selection and pretest is, in fact, a measure of reliability of the selection and pretest scores. It is equivalent to a test-retest reliability since the time interval between selection and pretest is assumed to be short. Thus, any appropriate reliability coefficient (e.g., internal consistency coefficient) could substitute for the correlation coefficient in the adjustment formula. This substitution makes the use of the adjustment procedure possible in cases where selection scores are used as the pretest measure and a separate pretest is not given.

Results of the analyses on absolute values (as displayed in Tables 4 and 5) provide some further insights. First, the similarity between the two sets of results in Tables 2 and 4 suggests that while the pretest mean score is not always higher than the selection mean score, exceptions are few. Furthermore, when the selection mean score does turn out to be higher than the pretest mean score, the mean difference tends to be relatively small. (This is confirmed by a perusal of the raw data.) Secondly, the dissimilarity between the two sets of results in Tables 3 and 5 indicates that while the adjustment procedure can, overall, be expected to work well with most data types, the accuracy of prediction is less than what the mean differences displayed in Table 3 would suggest. Furthermore, the accuracy of prediction appears to be a direct function of data reliability--as one would expect. Judging from the results in Table 5, it would appear that with data of relatively high reliability (e.g., data with error ratios of 9:1 and 8:2), the adjustment procedure could be depended upon to yield satisfactorily accurate predicted pretest mean scores, with absolute mean differences ranging from .45 to 1.56 for all the subsamples.

SUMMARY AND CONCLUSIONS

We have attempted in this investigation to assess the severity of regression effects when a set of selection scores is also used as pretest scores as this pertains to Model A of the Title I evaluation and reporting system. We have created data sets with various characteristics (i.e., data reliability and extremeness of subgroups) that are relevant to the regression phenomenon. These data sets were analyzed to obtain indices of the amount of regression which might occur under various conditions. An adjustment method was presented which predicted pretest scores on the basis of selection scores and the correlation coefficient between selection and pretest, the correlation coefficient being replaceable by such indices of reliability as the internal consistency coefficient.

Before discussing the implications of the results, it is necessary to point out some important limitations of the study. First, we have considered in this study only regression bias attributable to measurement error as it affects extreme subgroups. Other regression artifacts, such as those attributable to non-equivalency of control groups or differential growth rates for treatment and control groups (see, for example, Thorndike, 1942; Campbell & Erlebacher, 1970; Kenny, 1975; Campbell & Boruch, 1975; Bryk & Weisberg, 1977) were not examined. Consequently, while we would recommend the use of the adjustment method when its use is appropriate, it is obvious that such adjustment does not remove the other forms of regression bias which may occur with the use of Model A.

Secondly, we have studied only the impact of regression bias on the extreme low-scoring subgroups (which typically get selected to participate in Title I programs). Although one could quite confidently hypothesize that the impact of regression effects on the extreme high-scoring subgroups would, in terms of size, be similar to that

reported for the low-scoring subgroups, empirical testing of that hypothesis goes beyond the scope of the present study.

Overall, the results of this simulative study suggest that when data reliability is high (e.g., when reliability coefficients are .94 or above) the impact of regression effects appears small and perhaps negligible in most cases. The use of the adjustment procedure further reduces the size of the regression bias. This implies that when selection scores are obtained with an instrument of high reliability, such scores--with or even without adjustment--could be used as a pretest measure without severely biasing the evaluation results. This finding appears to hold regardless of how extreme the subgroup happens to be.

When data reliability is relatively low (e.g., when reliability coefficients fall between .84 and .70) the use of the adjustment procedure is likely to reduce the amount of bias to such a degree that the selection scores could still be used as a pretest measure without severely distorting the evaluation results. This is especially true of less extreme subgroups (e.g., subgroups formed by the bottom 20 or 30 percent of the group members) where the mean differences between pretest and predicted pretest are, in most cases, negligible. For subgroups formed by the bottom ten percent of the group members, the mean differences could, however, be rather substantial.

The findings of the study offer another option (besides using the adjustment method) to users of Model A. The results summarized in Tables 1-5 could be used to formulate a rule-of-thumb for adjusting selection scores when they are also to be used as pretest scores. More specifically, the mean differences reported for the various data types and extreme subgroups could be used as approximate indices of bias which had actually occurred in real data sets resembling the simulated data. These indices could then be used to adjust the means of selection scores accordingly. One could, for example, add the appropriate mean differences to selection mean scores and use the adjusted selection means as pretest

means. Alternatively, one could subtract the appropriate mean differences from mean gain scores when such gains have been computed with unadjusted selection scores serving as pretest scores. We would hasten to add, however, that the suggested rule-of-thumb should be used only as a last resort inasmuch as it yields only a crude and indirect estimate of the regression bias which might have occurred in the real data.

It should also be cautioned that data sets used in the present study have been created with normal distributions of selection and pretest scores. To the extent that a set of real data consists of selection or pretest scores that lack normality in distribution (such as when group members are rank ordered and those with lower ranks are selected to form the treatment group) the size of the regression bias may differ from those reported in this paper. Based on data derived from uniformly distributed random numbers, preliminary evidence suggests that the regression bias in those cases (i.e., when score distributions lack normality) will tend to be more substantial.

REFERENCES

Bryk, A. S., & Weisberg, H. I. Use of the nonequivalent control group design when subjects are growing. Psychological Bulletin, 1977, 84, 950-962.

Campbell, D. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), Compensatory education: A national debate, (Vol. 3). New York: Brunner/Mazel, 1970.

Campbell, D. T., & Boruch, R. F. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In A. Lumsdaine & C. Bennet (Eds.), Experiment and evaluation. New York: Academic Press, 1975.

Gulliksen, H. Theory of mental tests. New York: Wiley and Sons, 1950.

IBM. System/360 scientific subroutine package. New York: IBM, 1968.

Kenny, D. A. A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. Psychological Bulletin, 1975, 82, 345-362.

Tallmadge, G. K., The regression effect. ESEA Title I evaluation and reporting system Technical Paper No. 3. Mountain View, CA: RMC Research Corporation, 1976.

Tallmadge, G. K., & Wood, C. T. Users guide: ESEA Title I evaluation and reporting system. Mountain View, CA: RMC Research Corporation, 1976.

Thorndike, R. L. Regression fallacies in the matched groups experiment. Psychometrika, 1942, 7, 85-102.